Towards Ontology-Driven Approach for Data Warehouse Analysis

Case study : Healthcare domain

Lama El Sarraj[1,2], Bernard Espinasse[1]
[1]LSIS UMR 7296
Université d'Aix-Marseille,
Marseille, France
{firstname.lastname}@lsis.org

Thérèse Libourel[3]
[3]Espace-Dev UMR 228
Université Montpellier 2
Montpellier, France
therese.libourel@univ-montp2.fr

Sophie Rodier[2,]
[2]Assistance publique–Hôpitaux Marseille
DSIO
Marseille, France
{firstname.lastname}@AP-HM.fr

Abstract—**Understanding, reusing, and maintaining data warehouse resources is a key challenge for data warehouse users. Data warehouses resources are shared by different groups of users. The interpretation of information is subjective, it depends on user knowledge. Thus, a resource, like a data cube, is interpreted differently from a user to another. Unfortunately, misinterpreting data could induce serious problems and conflicts. To guarantee homogenous interpretation of data warehouse resources additional information is necessary. To tackle these challenges we propose to use ontologies to help the users in the exploitation of data warehouses. In this paper we propose an ontology-driven approach that represents data warehouse, dimensions and facts semantically enriched by their equivalent domain concepts and related to final resources provided by this data warehouse.**

Keywords*- data warehouse; ontology; decision information systems; decision making; healthcare institution management*

I. INTRODUCTION

Several surveys proved that big companies need efficient Decision support systems (DSS) and seek to expand the number of users over their DSS. To that aim, researchers found that companies need to have flexible decision tools, especially with, users' requirements and domain resources. A DSS is a collection of many tools or applications; we call them in this paper resources; that enable users to analyze, to query and to visualize a huge volume of data. In general, those data are stored in a data warehouse, and a set of Business Intelligence (BI) tools dedicated for data treatment and helping users (directors, managers, analysts, etc.) to make decisions.

Data Warehouse (DW) is the center of the DSS. DW is « a subject oriented, nonvolatile, integrated, time variant collection of data in support of management's decisions» [1]. In this paper we only consider resources provided by a data warehouse in a decision support system. To facilitate the task of DW analysis and treatment, a subset of the DW is created, it is called data mart. A data mart is oriented to a specific business need or a particular user requirement. Most of the times, data mart are organized in a multidimensional structure [2]. Data are represented like a point in a multidimensional space, visualized like a data cube (see Fig.1) [3]. They give users the possibility to synthetize and analyze data from three (or higher) dimensional array of values and various granularity levels. To manipulate data provided by the DW, end-users could use On Line Analytical Processing (OLAP) techniques, classic techniques, or even dashboards.

Taking user requirements into account is very important for the success or the failure of the DW [4], especially when users belong to different domains. The exploitation level of DW, as well as the preliminary conception level, is mainly based and adapted to user requirements [5]. Most research works devoted for DW focus on the approach design [6], [7], [8]. Even if these approaches are successful at the conceptual level knowledge about the data warehouse resources is still needed. It is important that users understand the semantic around the information he analyses and have a visibility about other resources that could help them to make efficient analysis.

The goal of this work is to design an ontology that relates data warehouse structure, resources and domain concepts. In consequence, in this paper we address two research questions:

- What are the competencies questions that our ontology takes in consideration?
- What are the concepts that compose the ontology to help decision makers in their analysis to understand indicators provided from a data warehouse?

Our research is supported by the public hospitals of Marseille; Assistance Publique Hôpitaux de Marseille (APHM). In this context we will present a case study from the healthcare domain specific to financial program based on the Program of Medicalization of Information Systems (PMSI) common to all French healthcare institutions.

This paper presents a new ontology-driven approach for DW personalization to resolute the semantic problematic related to the heterogeneous domains we applied our approach in healthcare management domain. The paper is organized as follow. Section II presents a case study from the healthcare domain. Section III presents the competencies questions that give an idea about the possible scenarios possible to help users in his analysis. Section IV presents the needed background. Section V presents an ontology-driven approach. Section VI presents an ontology-driven framework. Finally, before we conclude we present in section VII the related works.

ISBN: 978-1-61208-304-9

II. Case study

In this section we will present a case study from the healthcare domain specifically applied in the Program of Medicalization of Information Systems (PMSI). This case study is a good example that represents heterogeneous users that share same data warehouse.

In the French healthcare management system the PMSI has a central place. PMSI is a French adoption for the concept of Professor R. Fetter (Yale university, United States of America) to finance hospitals. PMSI specify the cost of sojourn based on diagnosis related groups that classes the hospitalization of patients in homogeneous and coherent medico-economic groups. This concept is applied in several countries like United States of America, England, etc.

In the healthcare domain users belong to the medical domain (doctors, pharmacists, biologists, etc.) whereas others don't (financial affaire managers, computer scientists, human resources, etc.). We should note that our approach is not limited to the healthcare domain. It could be applied in other business contexts where users are from different domains. This is, in general, the case of big institutions.

In this context we will take the example of a data warehouse. Fig.1 represents a data warehouse conceptual model for "PMSI activity" analysis. This DW conceptual model is composed of a fact table, dimensions, and measures.

Fact table = {Activity_PMSI}

Dimensions = {Date, Structure, Age, Exit_Mode, International_classification_of_desieases, Diagnosis_related_groups }

Measures = {Number of patient, …}

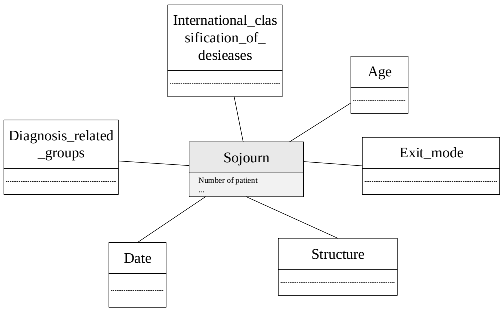


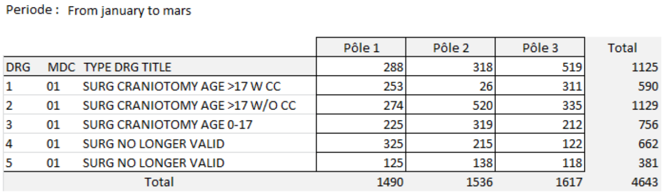
Figure 1. PMSI activity data warehouse conceptual model.

The multidimensional table (MT), MT = (M, D), where M is a set of measure and D is a set of dimensions. We will take an example of a multidimensional pivot table, presented in Fig. 2, for ethics reason we have taken fictive data:

D_1 = "Structure " (dimension level "pôle")

D_2 = "Diagnosis Related Groups" (attributes: DRG, MCD, TYPE DRG TITLE)

M_1 = "number of patients" (calculated measures: total of M1 per Diagnosis Related Groups, total of M1 per pole, total of M1 for all DRG and poles.



Periode : From january to mars

			Pôle 1	Pôle 2	Pôle 3	Total
DRG	MDC	TYPE DRG TITLE	288	318	519	1125
1	01	SURG CRANIOTOMY AGE >17 W CC	253	26	311	590
2	01	SURG CRANIOTOMY AGE >17 W/O CC	274	520	335	1129
3	01	SURG CRANIOTOMY AGE 0-17	225	319	212	756
4	01	SURG NO LONGER VALID	325	215	122	662
5	01	SURG NO LONGER VALID	125	138	118	381
		Total	1490	1536	1617	4643

Figure 2. PMSI pivot table.

In this research work we will take into consideration resources based on data warehouses sources and that represent data in a multidimensional table (defined by of measure, an operations on the measure, two or three dimensions, and a filter). In this context we noticed many difficulties:

Semantic lack

Users don't interpret the results in the same way. They need information about:

- Data warehouse concepts: dimensions definition, measures calculation methods and their sources
- Requirements expression heterogeneity: users don't belong to the same domain. They don't express their need with the same terms. For example: number of sojourn could be expressed as number of venue

Analysis needs

Most of the times, users need to analyze many resources to take a decision. In big institutions the big number of resources makes this task complicated. To facilitate this task, users need a global vision about the existing analysis axes. Thus, users need to have a global vision about the data warehouse structure to visualize the possibilities or existing resources that could help him to take a decision.

Finally, these difficulties lead us to think about a new semantic approach that structure the concepts related to the data warehouse based on ontologies.

III. Competencies question

In this section we exemplify and define possible scenarios to interrogate our ontology.

Entry 1: Data warehouse concept.

Output:

1. *Related data warehouse concept* -- Measures analysis -- What are the different measures related to an analysis axe? What is the different analysis axes related to a measure?
   Dimensions (Analysis axes) -- What are the measures that could be analyzed over a dimension?
2. *Resources concept* -- What are the existing resources to analyze a measure?
3. *Domain concepts* -- What are the existing measures to analyze a domain concept?

Entry 2: Resources concept.
Output:

1. *Data warehouse structure concepts* -- Which is the data warehouse (data mart) that provides a resource
2. *Domain concepts* -- What are the existing resources to analyze a domain concept?

Entry 3: Domain concept.
Output:

1. *Data warehouse structure* -- Which is the data warehouse (data mart) related to this domain concept?
2. *Resources concept* -- What are the resources to analyze a domain concept?

Those scenarios could be treated by using ontology technologies to visualize and have semantic to facilitate the analysis.

IV. BACKGROUND

In this section we will define the ontology and present some researches that have used ontology for the multidimensional systems.

A. *Ontologies*

Ontology is an explicit specification of shared conceptualization [9]. Different ontologies are proposed to define ontologies. W3C consortium recommends Ontology Web Language (OWL) to define ontologies. This language is based on the description Logic (DL) [10], it gives the opportunity to reason and represent structured knowledge. The DL language represents knowledge with concepts and roles. The concepts described as a set of individuals (instances) and roles describing a binary relation between individuals.

A knowledge base is represented with an ABOX (assertion box) and a TBOX (terminological box). An ABOX represent extensional knowledge (instances), TBOX describes the intentional knowledge of the domain as axioms.

We present the ontology with 4-uplet <C, P, ClassPropt, ClassAssoc> that concerns the TBOX.

Our ontology describes concepts to relate domain, resources and data warehouse structure. We consider:

- C represents the classes of the ontological model
- P represents the properties of the ontological model. P is partitioned into :
  - P_{value} : represents the characteristics properties
  - P_{fct} : represents domain dependent properties
- ClassPropt : C -> 2P relates each class to its property
- ClassAssoc : C -> (Opr, Expr (C)) is an expression that associate to each class an operator (inclusion or exclusion) and an expression to other classes.

B. *Multidimensional system*

We consider that DW resources are multidimensional table that represent a slice of the cube. The DW ontology registers the DW conceptual schema and the resources provided from this DW. For other purposes, several researchers like Prat et al [11] represents a multidimensional model with an OWL-DL ontology model, based on description logic [12], and define the transformation rules from the multidimensional level into OWL-DL ontologies. We will use these transformation rules to generate an OWL ontology of the DW model, based on transformations rules proposed in the work of Prat et al [11].

V. ONTOLOGY-DRIVEN APPROACH FOR DATA WAREHOUSE ANALYSIS

In this section we briefly present our approach and the architecture of our system.

Our approach focuses on two key requirements to address the research problem:

- It represents ontology architecture to describe knowledge about decision support system
- It provides an ontology-driven approach to help users in their analysis

A. *Approach architecture*

Our functional architecture Fig. 3 is based on three inter-related concepts, in order:

- Domain concepts
- Data warehouse structure
- Resources

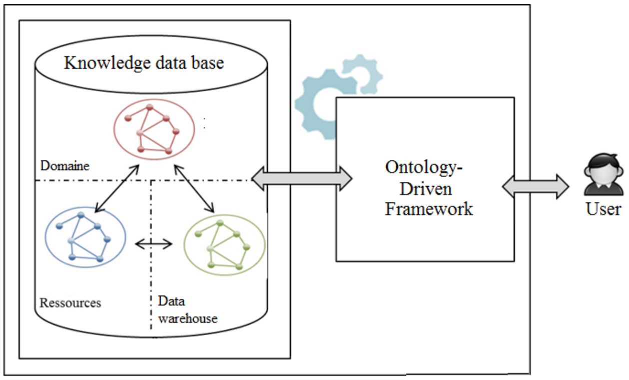


Figure 3. Approach architecture.

The framework system that we propose is based on an ontology interrelating three concepts (domain, DW and resources) to help users in the analysis task.

B. *Ontology concepts*

We will define the three concepts that compose our ontology. These concepts are necessary to help users in the analysis process:

Domain concepts structure: presents concepts of the domain and the relation between them. A decision is based on one or many indicators. In the analysis processes the user check the information's that he already know. However, most of the times user needs additional indicators to make his analysis. The domain description wills provide the information about the relation between domain concepts.

Data warehouse structure: the multidimensional model associated to the data warehouse organizes data into facts and dimension. Facts represent the subject of analysis and dimensions represent the axis of analysis. Fact table is the center of the multidimensional model. It stores elementary indicators, called measures. Dimensions can form hierarchies, structured in different granularity levels.

Resources structure: resources are provided by the data warehouse. Resources regroup information necessary for the analysis. To understand a component information about the indicator are needed like: calculation method, unit of measure, calculation period, date of creation, date of update, date of validity, objective, definition and the relation with the data mart.

C. Ontology connection

To connect those three concepts we will follow four steps:

1. Define domain ontology or use an existing domain ontology
2. Generate the data warehouse structure ontology based on the transformation rules proposed in the work of Prat et al [11].
3. Associate the data warehouse structure to the domain ontology, this step could be accomplished in several methods, for example :
   - Administrator relates data warehouse concepts to the domain concepts
   - Automatically align the data warehouse structure ontology with the existing domain ontology
4. Associate to the data warehouse concepts existing resources Ontology architecture

D. Ontology architechture

We will formalize our ontology by the triple < O_{DW}, O_D, *Map*> where:

- O_D is the domain ontology which provides a schema about the domain
- O_{DW} is a data warehouse schema which describes the resources (DSS components) related to the data warehouse
- Map is the mapping between O_{DW} and O_D which establish the connection between domain concepts and the DSS components

This ontology can be used for many purposes with ontology-based software. In the first hand, to give a vision about the relation between DW, resources and domain concepts, in the other hand, to propose for users other related resources to accomplish his analysis, based on the relation of the three concepts the resources, the data warehouse concepts and the domain concepts. Fig. 4 presents the ontology architecture meta-model to implement the knowledge base of the framework.

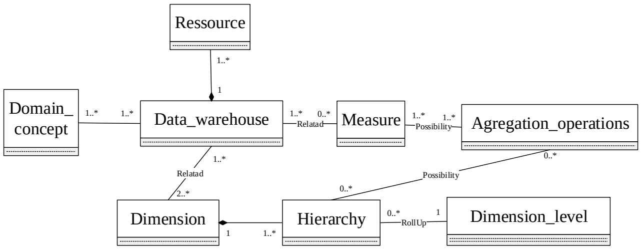


Figure 4. Ontology metamodel.

This ontology model represents the concepts related to the data warehouse. Each data warehouse is composed of zero or many measures and related to two or many dimensions. Hierarchies are composed of one or many dimensions. It is possible to effectuate operations on measures and aggregation according to the dimensions levels.

The proposed ontology model has been designed as follow to give high expressiveness about data warehouse components and to show the relation between DW concepts, resources (DSS components) and domain concepts.

VI. ONTOLOGY-DRIVEN FRAMEWORK

In this section we will present a framework based on our ontology. We implemented an ontology based on healthcare domain. Thus, this semantic structure will help users to discover and retrieve resources related to their domain and their first need.

To test our method we chose to implement OWL ontology with Protégé editor [13], and then we will use protégé to interrogate and visualize ontology with OntoGraph Fig. 5.

A. Methods

To create our OWL ontology we use “Protégé”, an open source Java tool providing an extensible architecture for the creation of customized knowledge-based applications.

1. Create three classes Data_Warehouse, Domain, and resources
2. Export existing domain ontology or create new domain ontology. These ontology concepts will be a subset of the domain class
3. Export data warehouse conceptual model ontology. To pass from the data warehouse conceptual model to OWL we applied the transformations rules proposed by [14]. Data warehouse concepts will be a subset of the Data_Warehouse class
4. Relate the data warehouse concepts to domain concepts. This task can be automatic by using existing ontology mapping tools; in this work we’ll not consider this option. To relate data warehouse concepts to domain concepts ontology administrator will refer to each data warehouse concept the equivalent, opposite, etc. concept in the domain ontology. For example, the data warehouse

dimension "Diagnosis_Related_Groups" will be related to "DRG" class of the domain ontology

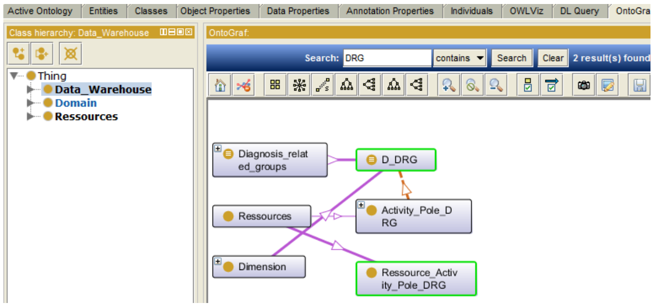
5. Relate the resources provided by the data warehouse to their corresponding concepts. For example, the resource named "PMSI_activity" allows user to analyze the PMSI activity per month and per medical unit. So, this resource will be related to Data_Warehouse subclasses dimensions month and medical units

B. *Visualization*

We will consider the example of the data warehouse presented in the healthcare domain. We will propose an ontology-driven framework.

Input: is a need expressed with a term or a group of terms.

Output: are concepts related to this need, about resources concepts, domain concepts, and data warehouse structure concepts.



Search: DRG contains

2 result(s) found

Figure 5. Example, retrieve 'DRG' concept from the ontology.

Thus, the user expresses his need with one or more keywords for example DRG.

- Domain concept: DRG is equivalent to "diagnosis related groups"
- DW concept: DRG is a dimension

So as Fig. 5 shows the resulting visualization of the ontology shows the existing concepts that contains DRG, equivalent and related concepts.

VII. Related Works

In the literature researches in the data warehousing field have already explored the ontology-based data warehouses and the personalization.

In the first hand, in the ontology-based data warehouses field researches are based on the multidimensional schema design, representation and its summarizability.

Prat et al [14] represent a multidimensional model with an OWL-DL ontology model to check the multidimensional model and its summarizability. Niemi and Niinimäki [15] provide an RDF model of an OLAP cube, they focus on the relationship between measure and dimension attributes and its effect on summarizability. They define the concept of measure-dimension consistency and they show how to conclude it from OLAP ontology. The OLAP ontology is constructed with semantic web technologies and is basically used to help users for OLAP cube construction and querying. Nebot et al [16] proposes a framework for designing semantic data warehouses. They propose the Semantic Data Warehouse to be a repository of ontologies and semantically annotated data resources and propose an ontology-driven framework to design multidimensional analysis models for Semantic Data Warehouses.

In the other hand, in the personalization of the data warehouse field we can distinguish three main objectives:

- *Customizing data sources schema* [17], [18] adapting the data structures to a specific needs of users
- *Customizing queries visualization* [19], or representation [20]
- *Recommendation of OLAP queries* [21, 22] to assist in the exploration of the ED

We also find the *personalization of the DW by recommendation that* can be associated to various works such as [17], [21], [23]-[26].

All these personalization techniques are not based on ontologies. Only Jerbi et al [27] adds semantic by annotation of the DW schema but his technique is not based on ontologies.

In our research we use ontology to personalize users need and retrieve not only semantic information about DW or cube schema but also the eventual existing resource like files (PDF, Excel, etc.), OLAP queries, etc. To that aim we integrate domain and resources concepts to our DW ontology.

VIII. Conclusion

The Data Warehouse (DW) resources are shared by users from heterogeneous domains. Those resources could be interpreted differently from a user to another. Consequently, semantic about those resources is necessary to guarantee the coherence of the analysis. Ontologies are effective solutions to add semantic to concepts. They facilitate the management of data, clarify and give a sense to ambiguous concepts.

Ontologies have been adopted by companies. Different solutions are offered to manage and query these data. In this paper we implemented the ontology with Protégé, interrogated and visualized the ontology with OntoGraph.

The study of concepts from healthcare domain confirms the need of semantic to help users in the analysis of resources provided by DW. One of the main characteristic of our proposed ontology architecture is that it provides a connection between domain concepts, data warehouse structure and data warehouse resources, this connection provide semantic information about resources and help users to choose other resources that can help him in his analysis. This personalization task is based on resources related to connected domain concept in the ontology.

ISBN: 978-1-61208-304-9

Furthermore, the main asset of our proposition is that it combines ontology and data warehouse to add semantic to resources analysis.

We should note that our approach is not restricted to the healthcare domain it could be applied for any domain for the retrieval of data warehouse resources.

This work leads to many other tasks. In future work, tasks that should be considered (i) test the integrity of the ontology when adding new concepts (like new resources), (ii) extension of this approach to add other type of resources and data source provided from decision support system but not related to the data warehouse, (iii) study different scenarios of the ontology evolution, (iv) validate our approach in a larger context.

Reference

[1] W. H. Inmon, Building the data warehouse, New York, NY, USA.: John Wiley & Sons, 1992.

[2] W. Lehner, “Modelling Large Scale OLAP Scenarios,” in Advances in Database Technology (EDBT), 1998, pp. 153-167.

[3] A. Bosworth, J. Gray, A. Layman , H. Pirahesh “Data Cube : A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total,” Data Min. Knowl. Discov., pp. 152-159, 1995.

[4] S. Rizzi, A. Abello, J. Lechtenborger, J. Trujillo "Research in data warehouse modeling and design: dead or alive?," Proceedings of the 9th ACM international workshop on Data warehousing and OLAP - DOLAP '06, pp. 3-10, 2006.

[5] M. Golfarelli, “From user requirements to conceptual design in data warehouse design – a survey,” 2009.

[6] R. Kimball, and M. Ross, The data warehousing toolkit, New York: John Wiley&Sons, 1996.

[7] N. Prat, and J. Akoka, “From UML to ROLAP multidimensional databases using a pivot model,” in 8èmes Journées Bases de Données Avancées, 2002, pp. 24.

[8] A. Tsois, N. Karayannidis, and T. K. Sellis, “Mac : Conceptual data modeling for olap,” in 3rd International Workshop on Design and Management of Data Warehouses (DMDW 2001), Theodoratos 2001, pp. 5.

[9] T. Gruber, “A translation approach to portable ontology specification,” Knowledge Acquisition, vol. 5, no. 2, pp. 199-220, 1993.

[10] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi “ The description logic handbook: theory, implementation, and applications,” Cambridge University Press 2003.

[11] N. Prat, J. Akoka, and I. Comyn-Wattiau, “Transforming multidimensional models into OWL-DL ontologies,” in RCIS, 2011.

[12] B. Grosof, I. Horrocks, R. Volz et al., “Description logic programs: combining logic programs with description logic,” in WWW, Budapest, Hungary, 2003.

[13] Stanford Center for Biomedical Informatics Research. 14/08/2013, 2013; http://protege.stanford.edu/.

[14] N. Prat, I. Megdiche, and J. Akoka, “Multidimensional Models Meet the Semantic Web: Defining and Reasoning on OWL-DL Ontologies for OLAP,” in DOLAP, Hawaii, USA, 2012.

[15] T. Niemi, and M. Niinimäki, “Ontologies and summarizability in OLAP,” in Proc. of SAC’10, Sierre, Switzerland, 2010.

[16] V. Nebot, R. Berlanga, J. Pérez, M. Aramburu, T. Pederson “Multidimensional integrated ontologies: a framework for designing semantic data warehouses,” Journal on Data Semantics, vol. XIII, 2009.

[17] F. Bentayeb, O. Boussaid, C. Favre, F. Ravat, O. Teste “Personnalisation dans les entrepôts de données : bilan et perspectives,” in Entrepôt de Données et Analyse en ligne (EDA), 2009.

[18] I. Garrigos, J. Pardillo, J.-N. Mazon, J. Trujillo., “A Conceptual Modeling Approach for OLAP Personalization,” in Conceptual Modeling-ER Verlag Berlin Heidelberg, 2009, pp. 401-414.

[19] L. Bellatreche, A. Giacometti, P. Marcel, H Mouloudi, D. Laurent “A personalization framework for OLAP queries,” in 8th International Workshop on Data Warehousing and OLAP, DOLAP’05, Bremen, Germany, 2005, pp. 9-18.

[20] D. Xin, J. Han, H. Cheng, X.,A. Li, “Answering top-k queries with multi-dimensional selections: The ranking cube approach,” in VLDB, 2006, pp. 463-475.

[21] A. Giacometti, P. Marcel, and E. Negre, "A Framework for Recommending OLAP Queries." pp. 73-80.

[22] A. Giacometti, P. Marcel, and E. Negre, “Recommending Multidimensional Queries ” in DaWaK, 2009, pp. 453-466.

[23] C. Sapia, “On Modeling and Predicting Query Behavior in OLAP Systems,” in DMDW, 1999, pp. 2.1-2.10.

[24] G. Chatzopoulou, M. Eirinaki, and N. Polyzotis, “Query Recommendations for Interactive Database Exploration,” in SSDBM, 2009, pp. 3-18.

[25] H. Jerbi, F. Ravat, O. Teste, G. Zurfluh, “Applying Recommendation Technology in OLAP Systems ” in ICEIS, 2009, pp. 220-233.

[26] A. Giacometti, P. Marcel, E. Negre, A. Soulet, "Query recommendations for OLAP discovery driven analysis." pp. 81-88.

[27] H. Jerbi, F. Ravat, O. Teste, G. Zurfluh, “Management of Context-Aware Preferences in Multidimensional Databases. ,” in ICDIM 2008, pp. 669-675.

ISBN: 978-1-61208-304-9